DOCUMENT RESUME

ED 151 930 EA 010 426

AUTHOR Mills, Richard P.
TITLE A Relational Model of the Financial Data Collected from Local Education Agencies by a State Department of Education.
PUB DATE Mar 78
NOTE 51p.; Paper presented at the Annual Meeting of the American Educational Research Asscciation (Toronto, Ontario, March 27-31, 1978); Figures 4, 5, 6, and 7 may be marginally legible due to small print

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Computer Science; *Data Bases; Data Collection; Educational Accountability; *Educational Finance; Elementary Secondary Education; *Information Systems; Models; *Programing; School Districts; *State Departments of Education

ABSTRACT

The educational community is recognizing that it has had little direct control over or knowledge of educational data collection. The objective of this paper is to present a description of presently collected financial data collected by a state education department. An entity set model is used to create a relational view of the data to facilitate development of an integrated data base. Methodology includes analysis of data collection instruments, interviews with users, and creation of relational models. The scope, structure, response burden, and adequacy of the financial data system are described. A neglected methodology is introduced to educational data description. (Author/MLF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED151930

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Richard P. Mills

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

A RELATIONAL MODEL OF THE FINANCIAL DATA COLLECTED FROM LOCAL EDUCATION AGENCIES BY A STATE DEPARTMENT OF EDUCATION

Richard P. Mills
Office of Economic and Fiscal Planning
Division of Research, Planning and Evaluation
New Jersey Department of Education

PRESENTED AT THE ANNUAL MEETING OF THE
AMERICAN EDUCATIONAL RESEARCH ASSOCIATION

TORONTO, MARCH, 1978

EA 010 426

A PROBLEM WITH EDUCATIONAL DATA MANAGEMENT

The collection of information from local education agencies has grown enormously in the last decade, to the point where many feel that the system is out of control. It is not always obvious that the data are of real use in state or federal planning and decision making. The response burden imposed by unsystematic data collection on local education agencies is costly and aggravating. Duplication of effort is such that several agencies or even several offices in the same agency collect virtually the same data. Much that is collected is untimely, inaccurate, belatedly reported, and therefore, useless. Much that is potentially useful is not in fact used. In plain words, the educational community is now recognizing that it has little direct control over or knowledge of what is presently collected.

This unfortunate situation is compounded by the nation-wide interest in accountability systems at the state level. Accountability systems impose new demands for data collection at a time when educators are all but overwhelmed by the information requirements of the old system.

At the federal level the data collection problem has gained the attention of the Council of Cheif State School Officers. In 1972 that body established a Committee on Evaluation and Information Systems (CEIS) and charged it with the development of a coordinated response to federal/state/local educational information systems problems.[1] CEIS has become an oversight panel to evaluate

[1] Council of Chief State School Officers, "Planning Educational Data Acquisition and Guidelines for Involvement of State and Local Education Agencies" (unpublished manuscript, Washington, D.C., August 1975).

new federal data collection plans and to give states adequate lead time to prevent federal intrusion into already overburdened state level data collection activities.

At the state level, the situation is exemplified by the new finance law in New Jersey, Chapter 212, Laws of 1975, which is similar to that of other states in that it establishes a statewide accountability system. Without systematic control of data collection, implementation of the "thorough and efficient system of public education" required by the new law will be difficult.

A delicate balance must be achieved. On the one hand, the scope of the monitoring and evaluating roles placed on the New Jersey Department of Education by court mandate and legislation is expected to evolve; this implies changing data gathering requirements. On the other hand, the control of data collection and revision of existing data collection instruments has been identified as a high priority concern by the State Board of Education.[2] Achievement of this important objective will require better information on what is now collected.

Response burden encompasses several measures of the amount of data being collected. Three measures used by the School Finance and Data Services Unit in the Colorado Department of Education were:[3]

1. The number of forms collected
2. The number of responses per form
3. The number of responses per year.

---

[2]New Jersey Administrative Code 6:8-6.1 (2).

[3]Charles M. Sisson, Dean Hirt, and Sharan Freden, Measuring the Response Burden of Data Acquisition (Denver: Colorado Department of Education, 1976), p. 2.

In the near future, Colorado will use two additional measures:

1. Hours required to complete each form
2. Cost of completing each form.

There are other measures that have been given less attention in the literature. For example, the response burden can also be examined from the perspective of the bureaucrat who has to process all that paper.

The frequency of data collection as measured by the submission date of the instruments is significant in a financial system because most financial reporting is concerned with specific reporting periods--end of month, end of quarter, end of year--unlike a personnel system that might require a status report as of an arbitrarily fixed date.

Another time-related measure of response burden is the effective date of the data items reported. The data item may be collected in September but is "true" as of the last day of June when the fiscal year ends. Knowledge of effective date is crucial when one is concerned with the order in which various changes in the data base are to be made.

Yet another measure of response burden is found by asking who the respondent is. Do all the forms land on the superintendent's desk or are others responsible? While it may be humane to spread out the response burden, it can result in chaotic data collection if two individuals are asked to supply the same data item by different agencies.

Response burden is a matter of concern here because it is a measure of what the future data system will have to accommodate.

Knowledge of this flow of data into the system in terms of the timing, who the respondents are, and many other factors must be part of the specifications of the data base and must be incorporated into a comprehensive model of a data collection system. There are pitfalls in the response burden approach, however. The approach nationwide to the data collection problem in education can be understood in terms of two objectives that are conflicting for the same limited resources of time, funds, and skilled analysts. The two objectives are (1) to reduce the number of forms, and (2) to create a data base. Figures on response burden add fuel to the short-term objective of reducing the number of forms. Unfortunately, forms control itself, while a vital first step, cannot resolve the long-term problems associated with the absence of an adequate data system. States that have developed impressive efforts in the data collection field report increases in response burden in spite of systematic planning.[4]

A theoretical issue underlying the long-range problem of defining an educational data base is how to specify the presently collected data, not just in terms of verbal definitions, but also in terms of other attributes that permit understanding of the logical relationships among data.

The problem addressed by this study was to describe the set of financial data items collected from LEAs by the New Jersey Department of Education and to conceptualize methods to verify and edit this description to facilitate the development of an integrated

---

[4]Michigan Department of Education, "Report of Department Data Collection, 1975-76 School Year" (Lansing, 1976), p. 3.

Sisson, Hirt, and Freden, Measuring the Response Burden of Data Acquisition.

data base.

REVIEW OF RELEVANT LITERATURE

The developing concept of a data base is a response to the generic problem of what data are needed in an organization, and how those data should be structured to be most useful. A data base is a set of related data stored without redundancy in a way that makes the data available for different applications while remaining independent of each application. The data are updated, modified, and used in accordance with a single set of management rules.[5]

The traditional data processing arrangement requires that the programmer and systems designer have detailed knowledge of all existing programs, each of which may be characterized by a unique logic, in order to recombine the data to resolve a new problem. Even when the data can be rearranged in a reasonable time period, this approach creates a new difficulty for the organization: the same data now exist in two or more places and the multiple files must be maintained and updated simultaneously whenever changes are made. As a result, the cost of maintaining accurate files and the probability of error increase. A data base approach, on the other hand, uses a single data structure that is flexible and thus applicable to all kinds of batch transactions without resorting. Independent data files are not established and so the redundancy problem is avoided.[6] The significant characteristics

[5] James Martin, Computer Data Base Organization (Englewood Cliffs, N.J.: Prentice-Hall, 1975), p. 19.

[6] G. Hubbard and N. Rover, "Automating Logical File Design," in Proceedings of the International Conference on Very Large Data Bases, Framingham, Mass., September 1975, p. 229.

of a computer data base are, in Martin's words, "non-redundancy," "data independence," and "interconnectedness."[7] Redundant data collection is to be avoided because it is expensive to update and maintain numerous entries of the same data.[8] Data should be collected once and stored in as few file locations as possible. Data independence means that the programs and the data structure are independent; changes can be made in the data structure without distorting the program, and vice versa.[9]

The first component of the data base problem, that of describing the organization's data, is made more difficult by the absence of a universally accepted terminology, a fact that is deplored by many students of data base design.[10] This confusion is due in part to the use of different approaches to the modeling of data bases, to the use of different terms by different manufacturers of computer hardware and software, and to the suddenness with which the problem of data description has emerged. The Committee on Data Systems Languages (CODASYL) labors to resolve these differences in the data processing field.

Models used for the description of data fall into two

---

[7] Martin, Computer Data Base Organization, p. 22.

[8] Ibid, p. 20.

[9] Ibid, p. 21.

[10] Beverly K. Kahn, "A Method for Describing Information Required by the Data Base Design Process" (Data translation Project, Graduate School of Business Administration and ISDOS Research Project, Department of Industrial and Operations Engineering, University of Michigan, 1976). p. 2.

categories: (1) the data structure set model and (2) the relational model. A data structure set is a group of related records referred to as members, which are organized, accessed, and controlled by one record designated as the owner.[11] An elementary example of such a model appears in figure 1.

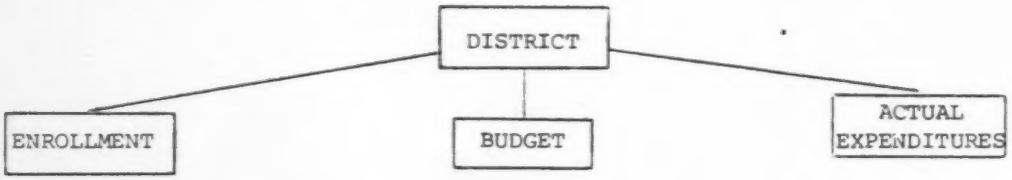


FIGURE 1 - A DATA STRUCTURE SET

Martin notes that the real difficulties in data base design stem from the requirement that relationships among data items must be defined.[12] A fully defined relationship conveys information that is essential to the design of a non-redundant data base; relationships provide the essential "connectedness" of a data base.

The literature reveals two conventions for mapping these data structure set relationships:

1. trees
2. plex structures

[11] Charles W. Bachman, "The Data Structure Set Model," in Proceedings of the 1974 ACM SIGMOD Workshop on Data Description, Access and Control, May 1974, p. 4.

[12] Martin, Computer Data Base Organization, p. 53.

A tree structure presents data in a hierarchy. The program structure model shown in the federal handbook series is an example of a tree structure. Using the conventional terminology, a tree is a hierarchy of elements called nodes, with one highest level node called the root.[13] The higher level nodes determine or define the lower level nodes; the higher node in each case is referred to as the "parent" and the related lower node is the "child." No child can have not than one parent, and therein lies an obvious problem with this method of presentation. Hierarchical representations are suitable in many modeling situations, but they compel the designer to seek out the one root or most significant record.[14] Many actual or potential users of data may not view the data structure in this manner.

An approach that overcomes this problem (while creating a few new ones) is the plex structure, or network, as shown in figure 2.

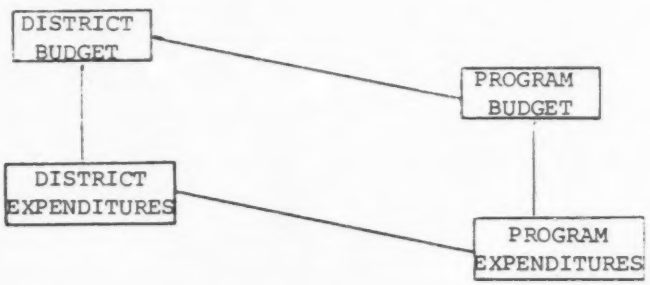


FIGURE 2 - A PLEX STRUCTURE

[13] Martin, Computer Data Base Organization, p. 79.

[14] John K. Lyon, The Data Base Administrator. (New York: John Wiley and Sons, 1976). p. 21.

Using a plex, designers can communicate about complicated structures by means of a notation composed of boxes to indicate records and arrows representing relationships between them.

Notations on the arrows are often used to indicate the kind of relationship that exists.[15] Relationships may be one to one, many to one, many to many, or one to many.

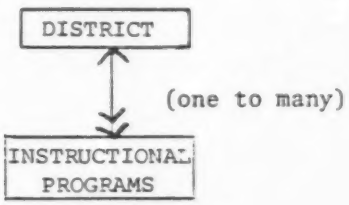


FIGURE 3 - MAPPING OF RELATIONSHIPS

A dotted arrow indicates a conditional relationship, one that may or may not exist in a particular instance. For example, a particular record "District" may be conditionally related to the "Title IV Program"; not all districts have such programs. The probability of such a link should be stated in the model. In a sizeable collection of data, the number of records and the relationships among them become so large that a plex model would be unintelligible to data users and thus useless as a model.[16] Codd has pioneered a second approach called normalization, which is based on the notion that a tree or plex structure can be decomposed

[15] Martin, Computer Data Base Organization, p. 68.

[16] Gayla Rampe, "An Approach to Data Base Design" (White Plains, N.Y.: AT&T Long lines, 1975), p. 4.

into a series of simple relationships presented in tabular form.[17] Codd's approach, referred to as the relational model, is concerned with the logical structure of the data as opposed to the physical structure. The logical structure must be simple enough for persons not trained in data processing--Codd refers to them as the "casual data users"--and yet comprehensive enough to adequately model the data. He argues that the casual user is most familiar with tables and therefore all data should be presented in that manner.[18]

Codd's tables are called relations and are composed of rows called tuples and columns called domains. A tuple consists of a set of related data items; it is a record. A domain is analogous to a field in traditional data processing; it is a part of a record.

The power of the relational model stems from the mathematically precise definition of a table. Codd notes that a data base table has properties that allow each row and column in a table to be identified uniquely no matter how they are reshuffled. The user thus has simple structures to deal with and he requires only a few simple computer operands to manipulate them.[19]

Codd has developed a relational mathematics that permits a "casual user" to create new tables in a precise manner. Unlike more traditional data processing operands, these manipulations affect entire tables. The user can create his own private arrangement of a portion of the data base as needed.

---

[17] Martin, Computer Data Base Organization, p. 150.

[18] E.F. Codd, "Normalized Data Base Structure: A Brief Tutorial." San Jose, California: IBM Research Laboratory, 1975. Mimeographed.

[19] C.J. Date, "Relational Data Base Concepts," Datamation, April 1976, p. 51.

It should be noted that certain problems exist with the relational model, chief among them that it has not been implemented on a large scale. It is also possible to form linkages which are invalid, although invalid links can be protected against by the operating system.

Data base technology is already enormously complex and it is developing rapidly. Educators will soon have to appraise its value and perhaps make use of it, but it should be clear from the foregoing overview that difficult preliminary steps in the identification, description, and control of collected data must be taken before this is possible.

The data dictionary is one important methodology leading to the description of collected data. It is a way to define "characteristics rather than content" of a data base.[20] A data dictionary is a system that defines the meaning, source, and use of all data elements.[21] Further, the data dictionary should include statistics on the use of data elements in order to reduce redundancy, eliminate unused items, and assess the impact of proposed changes.[22]

Ellis and Swanson surveyed state education department data identification activities and found sixteen states with data dictionary projects.[23] In terms of methodology the only common

---

[20] Jon J. Cahill, "A Dictionary/Directory Method for Building a Common MIS Data Base," Journal of Systems Management, November 1970, p. 61.

[21] Lyon, Data Base Administrator, p. 40.

[22] Ibid, p. 48.

[23] Ed Ellis and James R. Swanson, "State Data Identification National Survey" (Washington, D.C.: Committee on Evaluation and Information Systems, Council of Chief State School Officers, February 1974), p. 6.

thread seems to be an attempt to use a typology that is consistent with the federal handbook series.

The federal handbooks represent one important thrust in the description of educaticnal data. That series established standard terminology and definitions and organized these terms into various typologies for curriculum, staff, pupils, facilities, finance, and others, and many of the handbooks are interrelated in definitions and organizing strucutre. The purpose of these handbooks is to introduce a greater degree of comparability in the reporting of data.[24] One application of a data dictionary method is to conceptualize an ideal data base by conducting a data needs assessment and then to define the criteria for inclusion of an item in the data base. Colorado has taken this approach. The more commonly used method is that outlined by analysts in the Washington State Department of Education, where the key activities were the detailed study of existing data collection instruments coupled with interviews of users of the instruments.[25] Unlike Colorado, most data element dictionaries begin with the existing data collection for several reasons. For one thing, information in valuable.

A second reason for starting with the existing system in the design of the data element dictionary is that the information system, bad as it may be, is already in place. It is easier to make successive approximations toward the goal of an improved system than to redesign from scratch. Most educational information systems

[24] Charles T. Roberts and Allan R. Lichtenberger, Financial Accounting; Classifications and Standard Terminology for Local and State School Systems, State Educational Records and Reports Series: Handbook II, Revised (Washington, D.C.: U.S. Department of Health, Education and Welfare, 1973), p. 1.

[25] Washington State Department of Education, Information System Study, Phase II Report: Information Requirements (Olympia: Washington State Department of Education, 1974), p. 1.

consist of manual files, and one must be careful to avoid the assumption that the systems designed for educational organizations will be completely computerized.

Kahn finds that data base designers ignore the information structure perspective which defines the manner in which system users relate all data items in the system to create information.[26] That is, there exists in the real world a set of relationships among the entities described by the data system and these real world relationships have to be defined in an early phase of the data base design. The term "entity" is used to designate any object, concept, person, or thing about which information is stored.[27] An English program, a school district, and a classroom could all be regarded as entities in data base terms.

Rampe makes a distinction between the relational view, the usage view, and the structural view in describing the logical configuration of the data base. The relational view is the underlying structure of the data itself. It presents the essential components and relationships in the data system. The relational view defines the entities that are of concern.[28] The usage view is the structure from the perspective of programmers, systems designers, and users united in an effort to manipulate data for some specific purpose.[29] The structural view integrates the usage view and the relational view.[30]

---

[26]Kahn, "A Method for Describing Information," p. 2.

[27]Martin, Computer Data Base Organization, p. 441.

[28]Rampe, "An Approach to Data Base Design," p. 1

[29]Ibid.

[30]Ibid.

Senko, Altman, Astrahan, and Fehder suggested that the then available methodologies for defining information structures were imprecise, and that this difficulty was a significant problem for researchers in this field.[31] Mealy and, later, Engles defined three levels or realms of thinking about data. Level one is the real world of things and ideas called entities.[32] Real world entities, a school district for example, have properties such as size, location, or enrollment that are of interest. Level two concerns information about the real world entities. Here one speaks of entity records containing information about entities that correspond to the real world properties of the entity. At the third level is the world of data where individual data items about entities are identified. Martin, in his discussion of the Engles and Mealy conceptualization, warns of the confusion that results when analysts and others are not clear about which level they are dealing with.[33]

The entity set model is best stated by Senko, Altman, Astrahan, and Fehder, and elsewhere by Martin. According to this model, real world entities are viewed logically in terms of entity names (John Smith is the entity name of a particular entity), and

---

[31] M.E. Sanko, E.B. Altman, M.M. Istrahan, and P.L. Felder, "Data Structures and Accessing in Data Base Systems," in The Technology of Data Base Management Systems, ed. Richard A. Bassler and Jimmie J. Logan (Alexandria, Va: College Readings, 1976).

[32] R.W. Engles, "A Tutorial on Data Base Organization," IBM Corporation Technical Report TR.00.2004 (Poughkeepsie, N.Y.: IBM, 1970); G.H. Mealy, "Another Look at Data," in Proceedings of the AFIPS 1968 Fall Joint Computer Conference, pp. 525-534.

[33] Martin, Computer Data Base Organization, p. 44.

we regard entities with similar characteristics as members of sets which are given entity set names.[34] ("Employee" might be the entity set name for the set of which John Smith is a member.) It follows from the entity set model that the place to start in building an information structure would be to name entity sets and then describe these entity sets in terms of their attributes, and finally in terms of data items that relate to those attributes.[35]

The concept of a formally defined relationship is also part of the entity set model. Patterson writes that relationships can be distinguished on three levels--logical, entity, and physical.[36] An entity relationship is a relationship between two entities with similar or different properties. Kahn's more general definition of a relationship includes links between entity sets, individual entities, or entity properties. According to her definition, a fully defined relationship must include the entities linked, the kind of link, and the direction of the link.[37] Rampe sees a relationship as a set of directions leading from one entity to another.[38]

The differences between the data structure set model, the relational model, and the entity set model tend to become

---

[34]Senko et al., "Data Structures," p. 200.

[35]Ibid., p. 202.

[36]Patterson, "Requirements for a Generalized Data Base Management System," p. 108.

[37]Kahn, "A Method for Describing Information," p. 2.

[38]Rampe, "An Approach to Data Base Design," p. 3.

less distinct as they are applied by different authors who have grappled with the data base problem. In Bachman's application of a data structure set model, he adapts several entity set model concepts when he equates entities with records and attributes with fields within records.[39] Lyon appears to do the same in his application of what is essentially an entity set approach, to which he adds data structure set concepts. He uses a notation composed of boxes to indicate entities or attributes and arrows representing relationships between them.[40] In Lyon's model relationships represent membership or dependence.[41] In more precise language, the relationship represents a "set class," which is a category of identical linkages between identical entities or entity types.[42] The direction of the arrow representing the relationship is an essential feature of Lyon's structural models because it indicates the member of dependent entity, the entity whose existence is dependent upon the existence of the owner entity.[43]

It would perhaps be true, as Bachman has suggested, that the differences in these models are largely matters of style,[44] except for the very important distinctions between real world, information, and data that are at the core of the entity set model.

---

[39] Charles W. Bachman, "The Data Structure Set Model," In Proceedings of the 1974 ACM SIGMOD Workshop on Data Description, Access and Control, May 1974, p. 3.

[40] Lyon, Data Base Administrator, p. 147.

[41] Ibid., p. 148.

[42] Ibid., p. 150.

[43] Ibid., p. 159.

[44] Bachman, "The Data Structure Set Model," p. 10.

This difference is emphasized when one turns to the question of how to describe data in a step-by-step manner.

The preferred method is based on the notion that the firm owes its structure and form to the data flow.[45] Analysts identify the data at their source, determine how the data are used, and seek to eliminate multiple collections of data while finding multiple uses for the data. A key task of this approach is for analysts to continually validate their perception of the data flow and use in the organization by consultation with those who use the data. The data base structure must reflect the information structure used by the organization in achieving its primary objectives.[46]

The value of the entity set model is that it focuses attention on the real world entities that are of concern to the organization, and this emphasis compels a description and structuring of the data in terms of those real world concerns.

Rampe demonstrates this approach in her five-step methodology for the creation of a relational view of data:[47]

1. Name the entity/property sets in the system, including a definition of the relationships among them.

2. Eliminate duplication. Duplication includes data items derived from other items or sets.

3. Define relationship sets.

---

[45] J.D. Aron, "Information Systems in Perspective." In The Technology of Data Base Management Systems. Edited by Richard A. Bassler and Jimmie J. Logan. (Alexandria, Virginia: College Readings, 1976), p. 141.

[46] Richard F. Schubert, "Basic Concepts in Data Base Management Systems." Datamation, July 1972, p. 43.

[47] Rampe, "An Approach to Data Base Design," p. 4.

4. Document the relational view from the point of view of a single system. This is done for each system within the organization.

5. The individual system views are combined to form a "common view" which integrates all individual relational views.

It appears that the data description effort must consider all three of the entity set model realms in an iterative manner, but the effort should start from a macro view of the system. Unfortunately, there is often a rush to get to the inventory of data at the most detailed level.[48] Lyon on the other hand urges that the written description of the data should be presented to be criticized, modified, and validated by the user. He concludes that his macro level approach is not more time-consuming as critics charge; rather it focuses effort on the key area of logical design instead of on the less important business of coding.[49]

In education, the federal handbook series tends to encourage the overly detailed approach. These handbooks consist of exhaustive typologies of data or, in the case of Handbook II, Revised, of a chart of accounts for classifying financial data. That handbook points out that

> *these classifications as they are defined and coded, do not constitute a system. These items are grouped into mutually exclusive categories, sets, or dimensions...The procedure for filing and retrieving these dimensions becomes the system.*[50]

The dimensions necessary in a particular organization are to be selected from the handbook and organized as needed into an information system. All too often the approach is to use the federal handbooks as checklists to identify what data items exist.

---

[48] Hubbard and Rober, "Automating Logical File Design."

[49] Lyon, Data Base Administrator, p. 35.

[50] Roberts and Lichtenberger, Financial Accounting, p. 3.

RESEARCH DESIGN

The approach to the research problem relied on the entity set model and the distinction between the three levels of data description identified by Engles, Mealy, Martin, and others. This approach seems most useful in view of the problems encountered in previous educational data description efforts. It was also adopted because of the perceived need to verify the data description with data users; that verification requirement means that the data description produced must be simple and clear. The specific steps taken in the research procedures are based on those outlined by Lyon, Rampe, and Kahn.

The data collection instruments used by the New Jersey Department of Education as of March 1976 were identified, and a questionnaire collected data on the legal justification of the instrument, the respondent, the cost center collecting the data, the purpose of the data collection, and the name of the person responsible for each instrument. This person was identified as the system user for the purposes of this study.

Instruments that collected financial data were isolated from the above sample of instruments and were analyzed in terms of their entities and attributes. A series of entity set models was created based on the findings of the analysis, and these models were verified by consultation with users. The models were then consolidated into a department wide relational view of the financial data collected from local districts by the New Jersey Department of Education.

An instrument was said to collect financial data if:

(a) the instrument collected data items that are the names of either balance sheet accounts, revenues (to the district), expenditures, or cost measures; and (b) the instrument contains any other data items that relate directly to the financial operation of districts, schools, programs, or other local education agencies.

All approved financial instruments were inspected and the following operational definition of entity was applied:

> What is the concept, person, organizational unit, or other object about which information is being collected?

The answers to that question were tentatively identified as entities. All entities were listed by instrument and cost center and were circulated to division data analysts. It was assumed that the list at this point was incomplete and that additional entities would be identified as the analysis proceeded. This list of entities was much larger than expected, and this underscored the problem of relying on the federal handbook series at the state educational department level. Many of the entities were the result of the unique context of New Jersey statutes, and while these entities could have been forced into the handbook categories, the result would not be an adequate model of the data collection system because the relationships among entities within categories would not be clear.

The lists of entities were used to reexamine the instruments to list the attributes associated with each entity. The entity-attribute sets were studied to determine if items listed as attributes were actually entities. When an attribute defines or clarifies another attribute, that second attribute is actually an entity. These newly discovered entities were then listed separately

along with their attributes.

The entity-attribute sets were then used to create a structural model or plex, based on Lyon's notation, that depicted the relationships among the entities. A separate model was drawn for each instrument. These models were then consolidated into division-wide models. An example of these models is presented Figure 4. The boxes in the models represent entity sets and the arrows designate a set class. The direction of the arrows indicates the dependent member entity set in the set class. The existence of the member entity depends on the existence of the owner entity.

The attributes are not included in these models because this would needlessly complicate the plex. Relationships that were suspected but not confirmed by the entity property sets were not included in the models, but notation was made for future reference when the models were presented to users for validation.

The consolidation of the individual models is a tedious process that is based on the search for attributes in each entity-attribute set that convey information about the structural aspect of data items.

Users were interviewed and asked the following questions:

a. Are the entities correctly stated?

b. Are there any missing entities?

c. Are the entities related as shown on the chart?

d. Are there any relationships missing?

e. How are the submodels related to other entities or submodels on the chart?

This interview was an iterative process. Incorrectly stated entities were corrected, missing entities were added, and incorrect relationships were revised. All changes were made on the user's copy of the model which then served as documentation of the interview.

The revisions suggested by the users were used to prepare improved division level plex models.

The revised plex models in turn defined the relationships that existed among entity sets. Each relationship was defined in terms of the entity sets in the set class established by the relationship, the probability of the existence of the relationship, and the type of mapping (one to many, etc.).

The information on response burden identified those instruments from among the sample that were received by all districts. That limited sample of instruments was examined to identify the degree of redundancy in the data collection. Redundant items were defined as items that appeared more than once on each form, items that were derived from other items on the form, and items that appeared on more than one form. The non-redundant items were tabulated.

FINDINGS AND PRESENTATION OF THE MODELS

The research procedures made available three distinct files of data. The first was the data gathered from users through the completion of the questionnaire and the identification of a list of data collection instruments. The second set resulted from the detailed analysis of the instruments themselves. A third body

FIGURE 4

Chart of Entities and Relationships: Cost Center 03

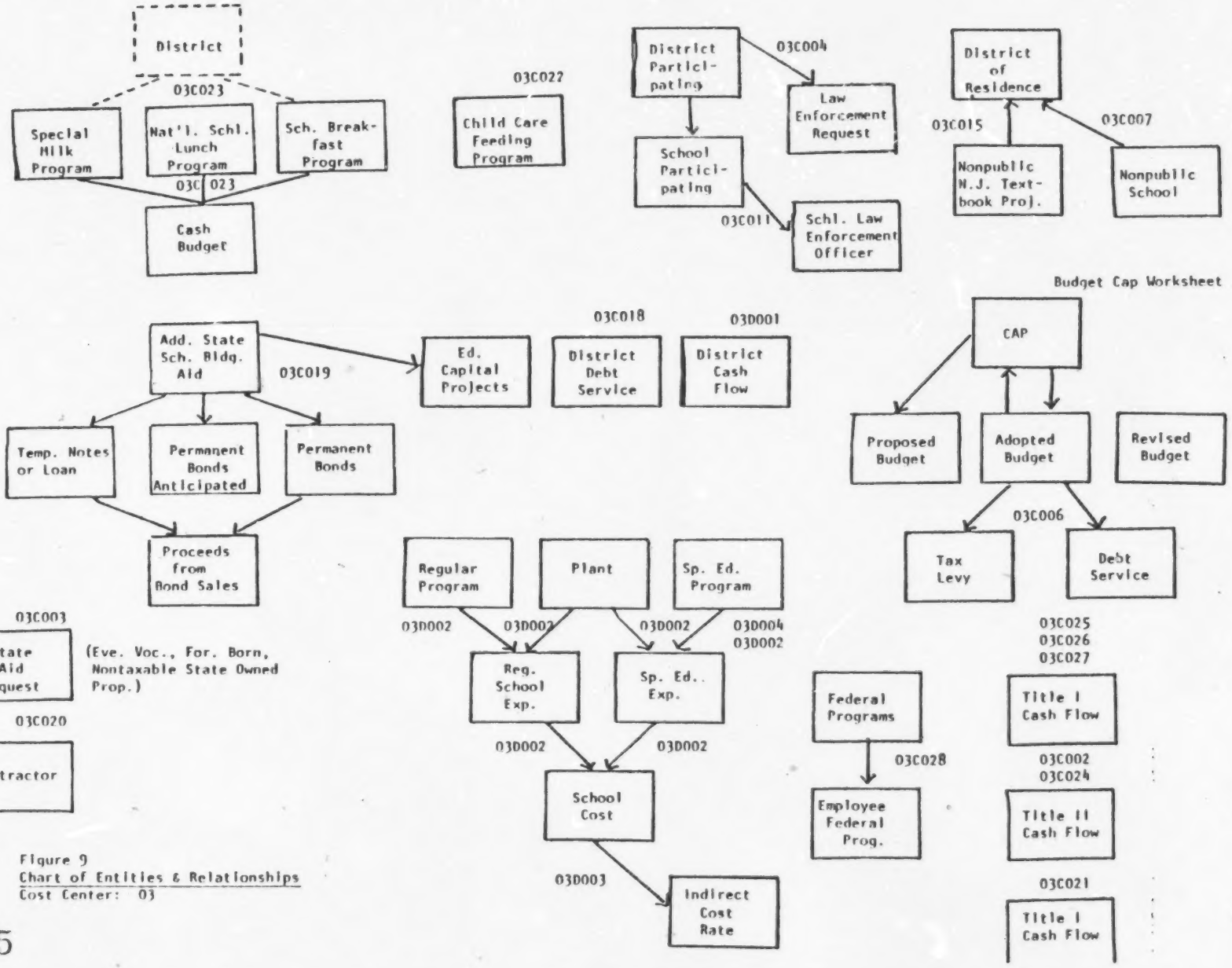


Figure 9
Chart of Entities & Relationships
Cost Center: 03

of data became available in the course of the consultations with users during which initial findings from the instrument analysis were presented in the form of a series of plex models and listings of entities and associated attributes that appeared on the financial data collection instruments. All three sets of data are needed in the description and organization's financial data collection activities.

A total of 57 instruments was finally identified as collectors of financial data. These instruments were the responsibility of five different divisions or cost centers in the Department of Education: Administration and Finance, Field Services, School Programs, Research, Planning and Evaluation, and Vocational Education.

Each of the five cost centers is unique in the way its data collection is justified. This finding suggests enormous difficulties in implementing an integrated data base. Data collection that is largely related to state needs may be much simpler to modify than that based on federal requirements. Although some states have been able to classify instruments as federal or state, it appears that in New Jersey a large proportion (39%) of all instruments collecting financial data carry both federal and state justification. Half of the instruments (29) are collected because of state law or regulation and a little more than one-third (22) are collected because of federal law or regulation. The collections of the different cost centers are distributed in different proportions among the federal-state categories. This finding suggests another difficulty in consolidating the financial data collected by the various cost

centers into a department wide data base; each division may give priority to a particular set of justifying regulations. The importance of state legislation as a justification for the data collected may explain Ellis' finding that each state education department data dictionary system tends to be unique. It also illustrates why a data identification process based exclusively on use of the federal handbook series as a checklist is likely to be unsuccessful: there are a great many items collected as a result of unique state law. It also suggests that data base design and the control of collected data may be difficult in a time when state school finance laws are changing rapidly as at present. A total of 34,535 instruments collecting financial data arrive at the five cost centers each year.

The study demonstrates that in broad outline the major periods in the updating of a financial data base would be the end of the fiscal year in June and the end of each quarter. December, in particular, has a heavy inflow of financial data, in part because of the budget calendar as established in Chapter 212, Laws of 1975, which requires that several reports be submitted for review in time for the school elections in the winter. If one examines each cost center in detail, it appears that this overall pattern is the result of the Administration and Finance collection; the other cost centers collect relatively more data at the end of the fiscal year, but otherwise their data collection activities are more evenly distributed throughout the year.

The date of collection is later than, and sometimes much later than, the effective date. The collection date is the deadline

for submitting the form and is thus a useful measure of the burden imposed on the respondent. The effective date is also a measure of the burden imposed on the receiver of the data and on the data system that must accommodate the data.

Different divisions at the state department level gather their data from different places in the district organization. The board secretary or business manager is the district person most often responsible for submitting financial data. What was not expected, however, is that an almost equal number of instruments are requested from program directors, which includes those in charge of federal programs, transportation, and other subsidiary units within an LEA. Often the program director gets the data he needs from the board secretary or business manager, but it is quite likely that the program manager keeps separate records for his one area of concern. This practice will result in different and conflicting data reports. It will be difficult in that respect to consolidate financial data when some of the elements come from one source and some from another.

A DESCRIPTION OF THE SET OF FINANCIAL DATA

The full description of the set of financial data collected from districts was presented in five parts.

1. Financial Instruments

Part one consists of a list of the instruments that collect financial data items and aggregates as of September 1, 1976. There is a total of 57 instruments on the list.

2. Entity-Attribute Sets

A second part of the description takes the form of a listing of entity attribute sets. This listing was arranged by instrument number, entity name, and attribute name and includes 115 entities that have associated with them one or more financial attributes.

3. Five Plex Models

The financial data collected by each division were shown in a plex structure in terms of entities and relationships among them (Refer to Figure 5). Solid arrows denote relationships that must exist, given the existence of the owner entity. The arrow indicates the direction of the owner relationship. Dotted arrows indicate conditional relationships that may or may not exist in a given occurrence of the owner entity. Entities are indicated by solid boxes. Each entity has associated with it one or more attributes, which can be found by reference to the entity attribute sets above.

The series of division level models can be simplified in two ways. First, the models can be limited to those entities with associated financial attributes; this modification is, of course, based on the list of entity-attribute sets. A further improvement in the plex models results from a classification of entities into three types: organization entities, financial entities, and others. Schools, districts, programs, and projects relate to various organizational levels in an LEA and can be grouped together as organizational entities. Budgets, expenditures, cash flow, bonds, investments, and the like constitute a wholly different category, which was characterized as financial entities. The

FIGURE 5

Revised Chart of Entities and Relationships: Cost Center 03

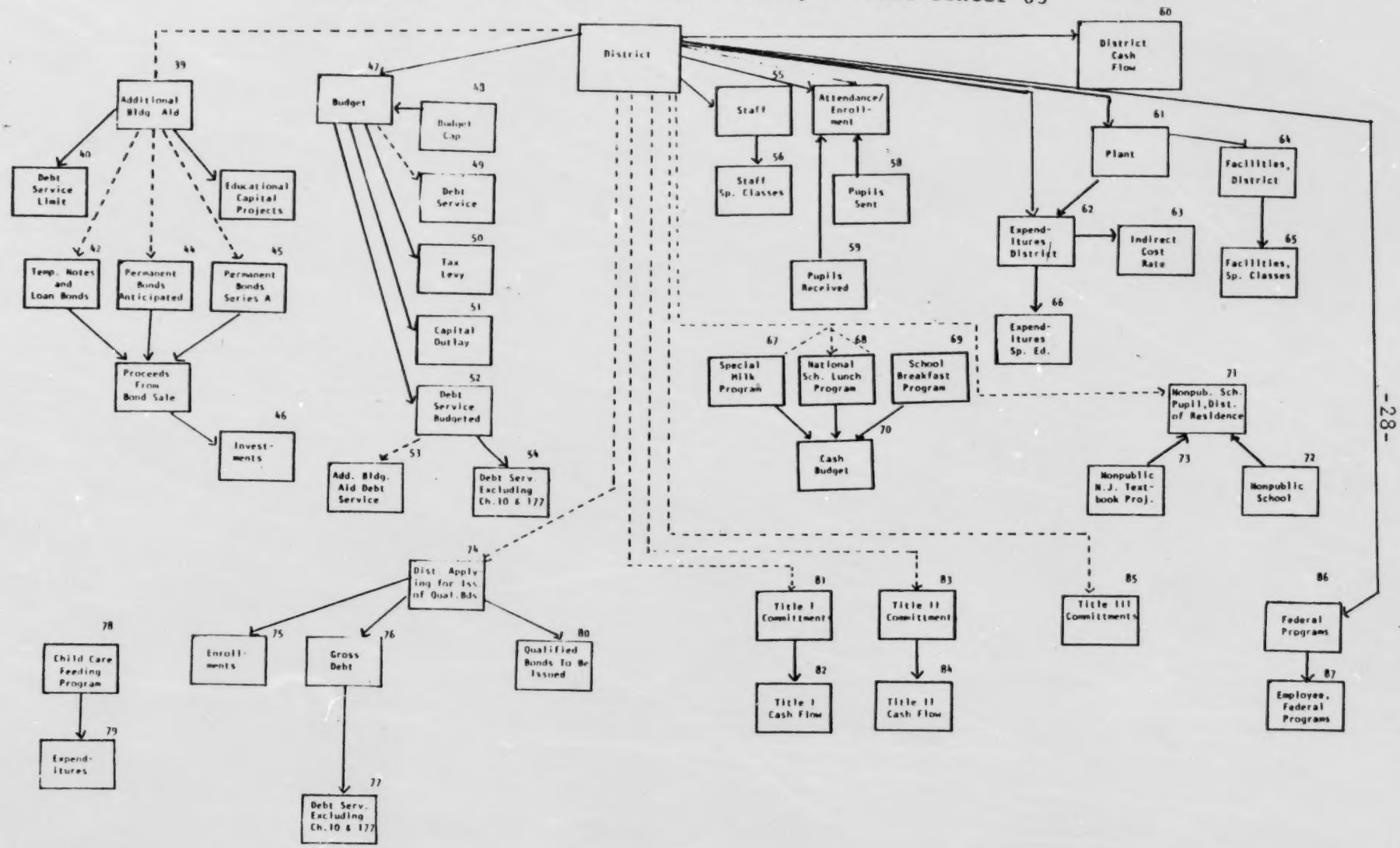


Figure 11

Revised Chart of Entities & Relationships

Cost Center: 03

remaining entities that fit neither of the two above groups appeared to be concepts such as "debt service limit" and these were grouped into a third category. A new set of plex models was drawn based on these two simplifying measures.

The division or cost center level structures have been defined only after considerable difficulty. The last remaining step, which is to achieve a department level consolidation, is now a relatively simple task. This department-wide model is shown in two steps in Figures 6 and 7. Figure 6 links all district-related organizational, financial, and other entities at a district level of aggregation. Figure 7 carries the analysis to a more detailed program project level of aggregation although, to avoid an overly complex representation, only the organizational entities are shown. Financial and other entities related to these organizational entities can be found by reference to the division plex models.

4. Relationships Between Entities

The fourth section includes an index of relationships between entities. A total of 95 unique relationships were identified by examination of the plex models that had been verified by the users and/or user representatives and simplified as described above. This index, established the structural aspect of the description of the financial data system. Each relationship is named in terms of the entities that are linked, with the owner entities listed first. Also shown is the probability of the existence of each relationship, given the existence of the owner entity and the type of relationship that exists, whether it be one to one, one to many, many to many, or many to one.

FIGURE 6

Consolidated Plex Model--District Level

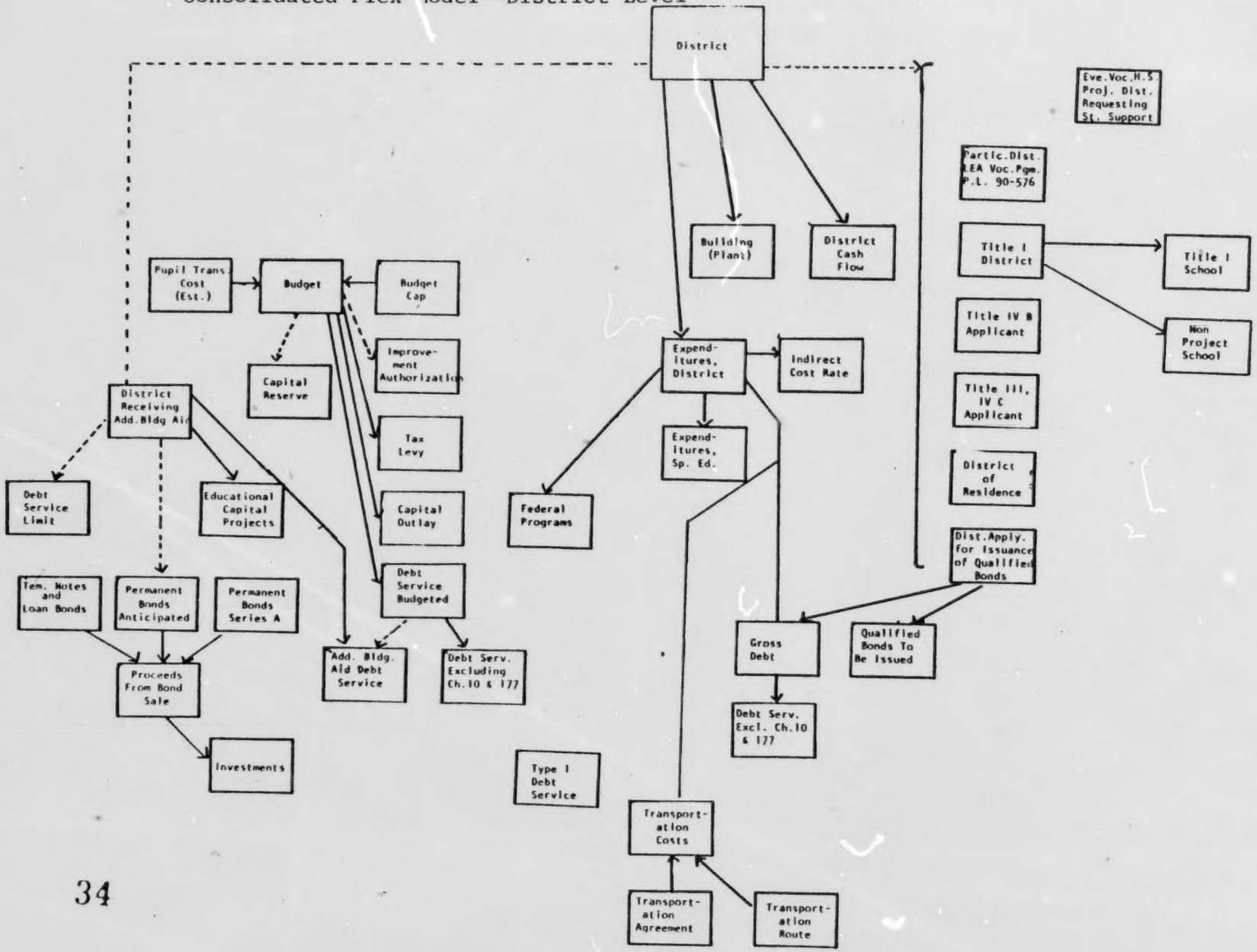
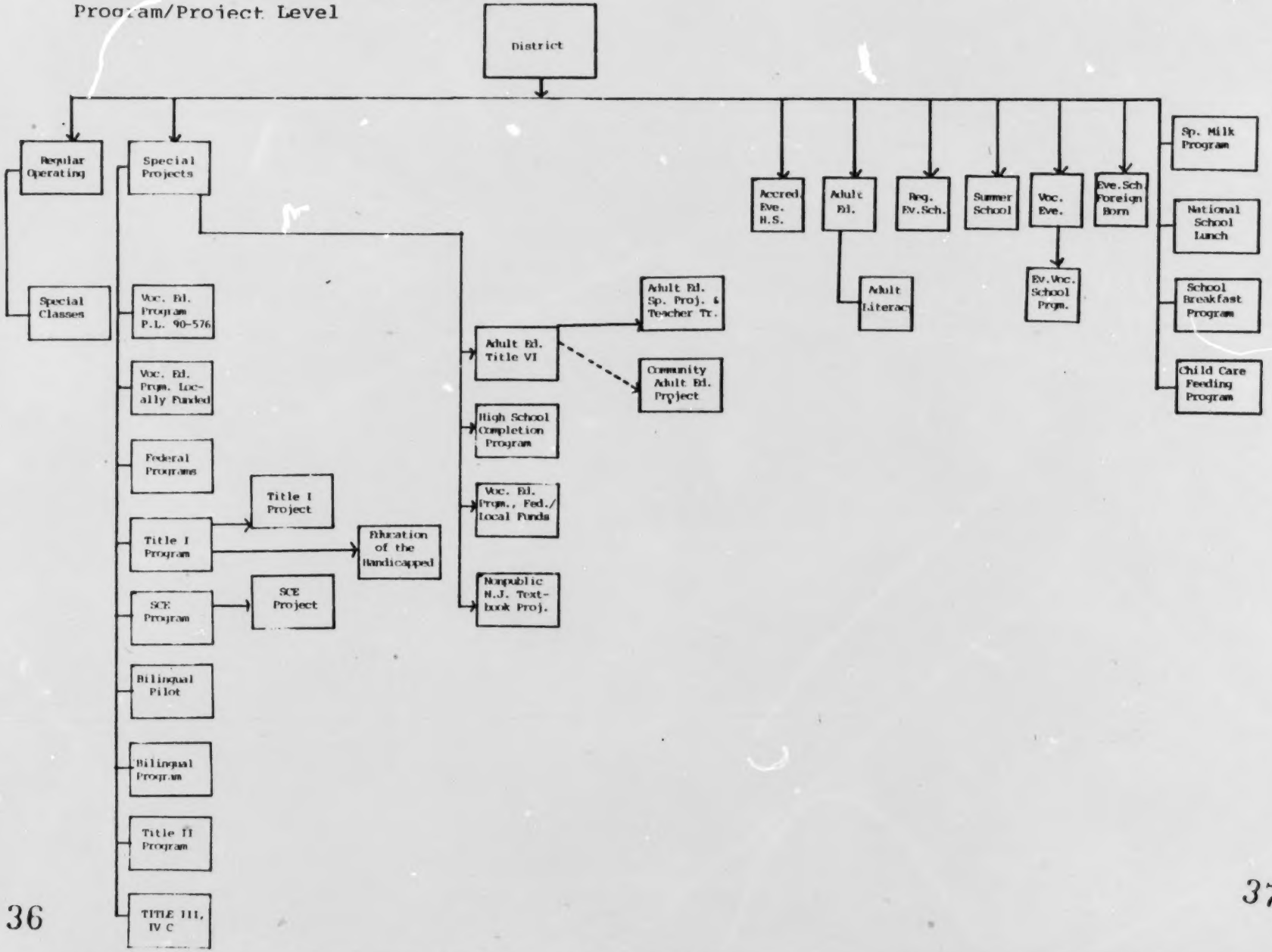


FIGURE 7

Consolidated Plex Model--
Program/Project Level



5. A Core of Financial Data

Yet another way to describe the financial data is to distinguish between a central core of data collected about every district and the data collected only about districts or other entities due to participation in educational programs. There are several reasons for describing the core of data available for every district. It can be argued that the first priority in terms of forms revision, reduction of duplicate data collection, and development of a computer data base would be such a core because it would affect the largest number of respondents, and such a project would have the widest possible use for further analysis. That core of data would in all probability be used for the most important standard reports.

It appears that only 12 instruments are collected from all districts, while the other 45 are due to participation in some special program or to some other atypical circumstance.

The financial data collected by the 12 instruments identified above were subjected to further analysis to describe the data items themselves in a data dictionary.

It was found that 661 non-duplicated data items should be considered as part of the core, and that 1,174 redundant items are collected on the instruments.

A total of 63.9 percent of the financial items collected on the core sets of instruments are redundant in the sense that they are derived from other items on the form or simply copied from other instruments.

The plex model in Figure 8 presents the structural view of the entity records and relationships that are collected by this

core of financial data collection instruments. It represents a consolidated, organization-wide model of the collected financial data available concerning all districts in the State of New Jersey.

DISCUSSION OF THE FINDINGS

Both the theoretical model used and the knowledge gained during the research procedure point to the data structure as the most significant aspect of the data collection.

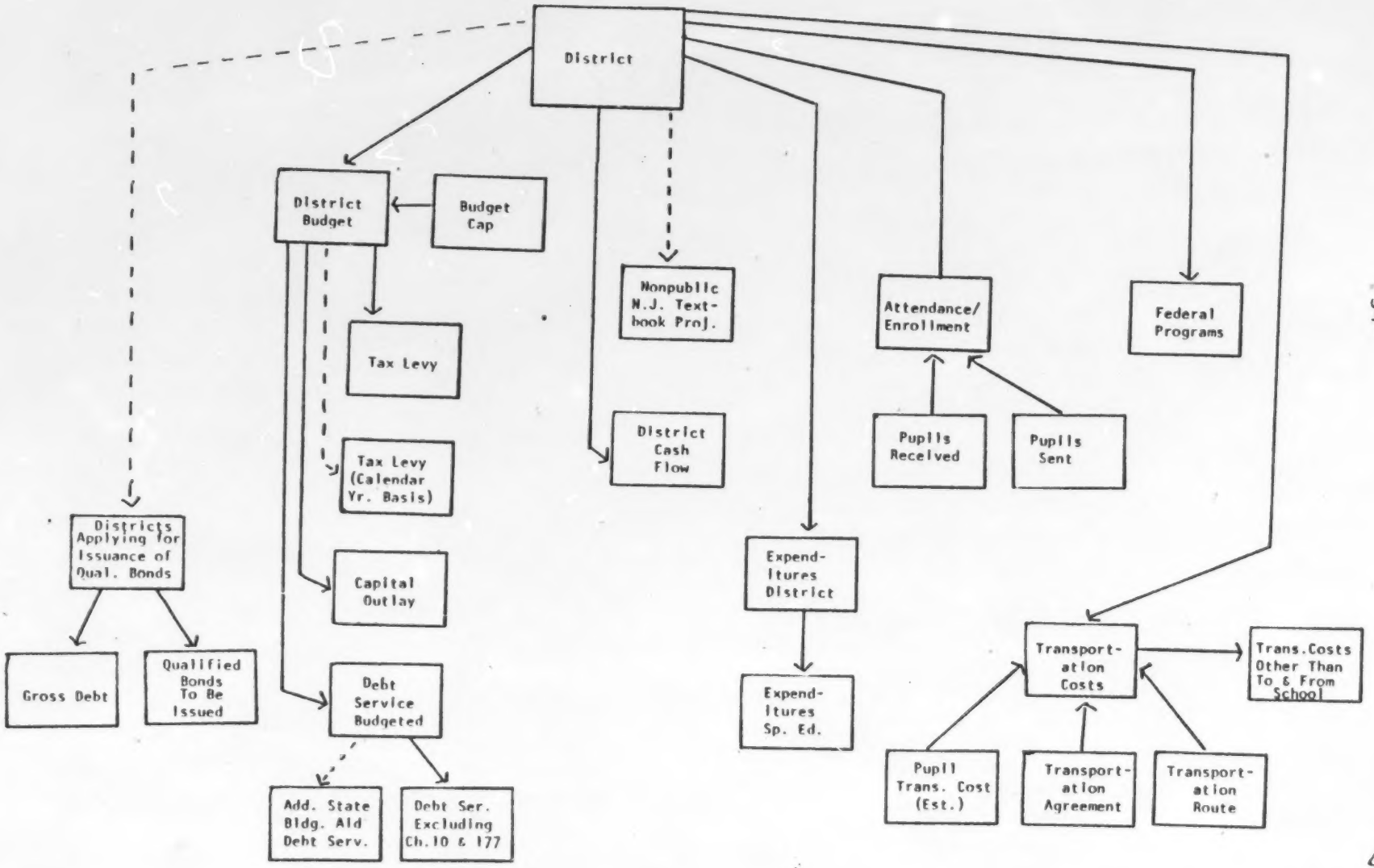
The entity records and relationships that exist in the case of every district are quite small, as shown by the final plex model. This finding suggests that the financial data base be conceptualized as several distinct data bases:

1. Core data concerning all districts.
2. ESEA programs.
3. State compensatory education programs.
4. Bilingual programs.
5. Transportation.
6. CETA.
7. Vocational education.
8. Other unrelated state or federal programs, such as the school lunch program or the non-public textbook program.

This list appears to lead to acceptance of a fragmented data collection operation with each cost center possessing its own data. On the basis of both theoretical and practical considerations, this must be rejected. It may be possible, however, to define a

FIGURE 8

Entities and Relationships in the Core of Financial Data Collected from All Districts



distributed data base that consists of the core data base with pointers to a number of subsidiary data bases. Such a solution would retain the logical structure that relates all entity records and would also retain the separate identity of the subsidiary data bases. The distributed data base solution might also make implementation of the data base (i.e., adoption of common collection, coding, storage, and processing procedures) less disruptive because the various parts of the organization would still retain a coherent view of "their" data. Unfortunately, the acceptability of the distributed data base concept in this case cannot be evaluated without a further step: the creation of the usage view to complement the relational view presented in this paper. The usage view, in general, would define the extent to which the apparently separate structures of data are used together. In other words, how likely is it that a data user would have to draw data from more than one structure? A high degree of such use would argue for a high degree of integration among the structures.

Another policy issue concerns the question of the adequacy of the structure itself. That is, does the structure defined by the entity records themselves coincide with the structure of real world entities that should be of interest and in fact are of interest to those who actually provide education to students? In this context it has been found that the cost centers differ in the level of aggregation in the data collected. The Division of Administration and Finance is primarily interested in district wide data, while most other financial data are collected about schools. In short,

it is apparently not possible to query the set of financial data about the cost of educational services at the level at which they are delivered except in a few isolated cases.

This research has concluded that even the program level data that are actually collected deal with program entities that are outside the area of regular instruction; they are more likely to be vocational or special instructional programs. The plex models reveal, therefore, that questions involving financial data about regular instructional programs cannot be answered by the data presently collected.

The data structure also has implications for the new accountability measures being introduced in New Jersey school districts as a result of the New Jersey Supreme Court mandate and the enactment of N.J.S.A. 18A:7A, which will require districts to develop budgets in a program oriented manner. That is, districts must identify objectives based on district goals and needs, develop programs to achieve those objectives, and prepare budgets for those programs. In order to accomplish this, districts will have to develop unique program structures; there can be no program oriented approach to budget preparation without an explicit program structure. These district program structures will constitute data structures similar to those modeled at a state department level in this paper. It would not be surprising to find that it is difficult to induce local districts to use data for systematic planning, as is the intent of the new accountability law, since so much of their data collection and reporting is geared to responding to the state education department. The existence of an unplanned and, until now, undefined state department data structure may tend to force LEAs away from

what would be coherent structures for financial reporting systems at the local level.

It cannot be denied that each cost center plex model is based on planning, but the overall combination of these models is what individual districts perceive. It is this overall model that is unplanned.

The incoherent structure of the financial data structure is not wholly unexpected in view of the justification cited for the collection of each data instrument. Each instrument is justified by some state and/or federal law or regulation. The primary reason for the collection appears to be to serve a control function. Public finance has always emphasized the control aspect of budgets; budgets are, in the public sector, legal documents that put explicit limits on spending. The sheer volume of data collected for this important purpose may so overburden the capabilities of the local data system that there is little time for the cost analysis, cost projections, and many other uses of financial data at both a state and district level that should be considered in planning the data collection system. Is the data set presently collected adequate for these additional purposes? The answer based on this research is that it is probably not adequate because it is at such a high level of aggregation except in the areas of the specialized programs. As a result, it is generally not possible to relate costs to outcomes with any degree of precision in the regular instructional area.

The financial data structure, when examined from the point of view of a state department of education, is seen in terms of what is mandated. The financial data structure should

be designed so that what is required at the state and federal level is the result of an internal data collection and reporting system that is of real use to local districts who exist, after all, to teach students and not to report to higher levels of government.

The financial data collection has been implemented at an operational level as opposed to functional or corporate levels. There are a number of quasi-functional level systems within the organization but no overall financial function exists; different cost centers have different pieces of the financial picture. The findings of this study point to the need for a quantum leap from an operational level system to a corporate level one, without the intervening concentration on functional data systems. In other words, it is not sufficient to merely tie all the existing data together into a system similar to what has been depicted in the plex models. The system must be designed as a coherent structure based on the usage view, and it must serve overall policy needs as well as management and operational needs. The corporate data base concept must include all the levels below it. If a corporate level data base model were to be designed without consideration of the more numerous operations-oriented users, the data base would be out of touch with the suppliers of data, and none of the problems that motivated this inquiry would be resolved.

The findings on response burden support, at least in the area of financial data collection, the commonly held assumption that there is a high degree of duplication in data collection. The data collection is redundant in that a high percentage of the

data items collected are copied from other forms or from other pages on the same form and, further, that a great many items can be derived from other items. On the other hand, there is relatively little duplication in the sense that the same financial items are collected about the same entity records in more than one cost center.

What can be done to reduce the duplication of effort? The first and most obvious solution is to revise data collection instruments. The application of a common set of standards for forms design can eliminate much of the problem.

A more fundamental approach to the problem of redundancy is to shift thinking patterns about financial data collection. The data items are part of an organization-wide structure; while the structure itself remains relatively constant, the details are changing in a predictable manner as the values or items with particular effective dates are replaced by other values. The important consideration is to determine the uses made of these items. What are the tasks, decisions, and processes that make use of the items? Which tasks are routine and which are special, one of a kind data uses? It may well be that there are other important tasks that are overlooked because the existing structure of the data tends to focus questions into relatively narrow channels.

One final point is that the findings (pages 23-26) indicate that the financial data are not only redundant but also fragmented. Five divisions collect budget data. In many cases there is a relationship between these budgets but often there is not.

The budget data are collected in overlapping categories. One instrument collects vocational education budgets that include

state and local expenditures, while another collects only local expenditures. It would be surprising if there were not a high degree of confusion and inaccuracy about what is reported on these forms.

The different programs require financial data to be reported in many dimensions:

- grade
- grade cluster
- project area
- school
- type of student handicap
- instructional program
- object of expense
- source of funds
- category of expense (groups of object codes unique to a particular form)

This situation forces the district with any involvement in special programs to adopt a confusing, multi-dimensional accounting system to capture the data in the manner in which it must be reported.

EVALUATION OF THE RELATIONAL VIEW APPROACH

In the state education department environment, the relational view approach used to resolve the research problem appears to have been more successful than were the item analysis procedures first attempted.

The relational view approach is characterized by iterative attempts to model the data collection environment, while the item analysis that is favored by state education department data systems that were reviewed begins with an assumption that the structural model is intuitively obvious. This application of the relational view methodology demonstrates that the structural model can be delineated only with great difficulty and, further, that the resulting

model is quite unlike that presented in the federal handbook series. Failure to recognize this overall structure as it is actually defined by the data can lead to a premature concern with the details--the data item. If the structure is not identified and then initially examined in terms of the major objectives and purposes of the organization as a whole, it is probable that attempts to create a data base will result in a simple automation and speeding up of the unplanned activities that already exist in a bureaucracy.

The importance of the user to the development of a data base and even to the development of short-term improvements in the data collection system cannot be overemphasized. Without repeated contact with the user, the analysis can become detached from the reality. Unless a macro modeling approach such as the relational view is used instead of the item analysis approach, the results of the analysis will be too voluminous and too technical to be verified, edited, and ultimately improved by the user.

Another result of the study has been to demonstrate the analytical processes that must be completed before the effort becomes involved with the computer. For example, identification of the instruments themselves was a long and difficult procedure. Without forms control, a department of education can go no further. Once forms control is established, it must be maintained. Data collection needs appear to change so rapidly that even within the time period of this analysis, several instruments were added and deleted.

The data collection instrument has been viewed by this

study as an artifact of a data system that could be identified and specified by means of analysis of those instruments and consultation with users.

The question of what defines the boundaries of the system depends upon the objectives pursued in the analytical task. Plagman and Altschuler state that a data dictionary effort deals with two objectives:[51]

1. To present users with information about the collection and dissemination of data;

2. To establish standards for the naming and usage of data.

The first objective has been met by this descriptive study. The data are explicitly described in terms of entities, attributes, and relationships. The second objective cannot be achieved at this time for the reasons mentioned regarding the condition of the data collection instruments and, in any case, is beyond the scope of this inquiry.

Another way of looking at objectives is in the terms used above to describe the conflicting objectives of state educational data description efforts. The immediate need is to improve the control of collected data; that is, to reduce the response burden. From that point of view the study has correctly limited the boundaries to financial data collected from districts. From the point of view of the long-range objectives of the definition of a data base, these boundaries will have to be expanded to include all data of interest to those in and out of the department of education. There are

[51] B.K. Plagman and G.P. Altschuler, "A Data Dictionary/Directory System within the Context of an Integrated Corporate Data Base," in Proceedings of the AFIPS 1972 Fall Joint Computer Conference, p. 1136.

several other sources of data that are not accessible through the analysis of instruments.

The key concept in this paper has been the description of data in terms of the entity set model or the relational view. This is a neglected view of educational data; most educational data studies have examined only the detailed items. Data collection reform can be achieved by standardizing terminology and reporting standards at the item level, but major advances will probably await an overall structureal design at the entity record level.

The relational view has been applied to the limited field of financial data only. It can be extended to encompass all data collected by a department of education. The concept is useful because it allows gaps and logical inconsistencies in the data collection activity to be identified. It results in a much more concise statement of the data than the item level analysis generally employed, and for that reason it is more accessible to those who seek policy level information about how data collection and use can be improved. The relational model would also be useful as a general approach to be employed by other state education departments that are beginning the study and control of financial data. States that have already begun data analysis on the item level should at least consider the macro view of the data structure before attempting to computerize their existing data. Many states are moving in the direction of greater accountability for districts, which means that educational activities should be related to explicitly stated goals. State education departments are recognizing that they, too, must meet similar standards in carrying out their functions. As it happens, many of the functions of state education departments revolve around

the collection of valid and timely data. The relational approach should be considered because it helps an organization "see" its data in such a way that the data collection can be examined in terms of broad goals and objectives, and reshaped if needed.